
Myrinet[®]

Cluster Interconnect

Charles L. Seitz
CEO & CTO of Myricom, Inc.
chuck@myri.com

21 June 2005
ISC2005 Interconnect Tutorial

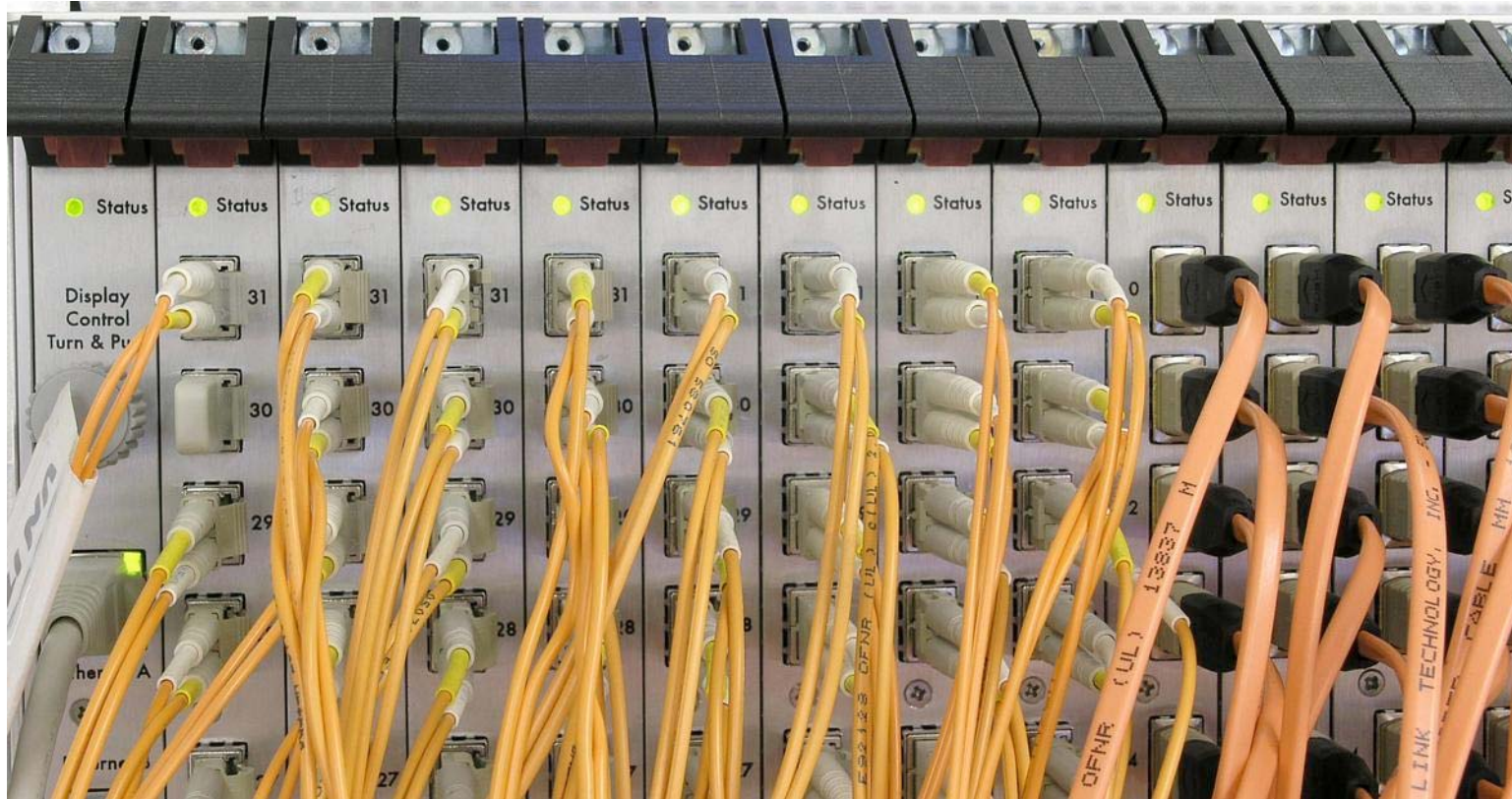
What is Myrinet?

- A **network architecture**, protocol, and technology
 - A descendant of packet communication and routing in MPPs, but open.
 - ANSI Standard (ANSI/VITA 26-1998).
 - **The mirror image of Ethernet:** Processing power is concentrated in the hosts and NICs, allowing a streamlined, highly scalable, switching technology.

However, it helps to understand Myrinet first as:

- A set of **standard products**
 - Network-interface cards (NICs), software, switches, and cables.
 - All you need to make a high-performance cluster from a set of host computers.
 - *Install the NICs and software in the hosts, and connect the NIC ports with cables and switches.*

Myrinet-2000 fiber links, 2+2 Gbit/s, full duplex



Advantages of fiber: small-diameter, lightweight, flexible cables; reliability; EMC; 200m length; small connectors; low-cost industry-standard 50/125 multimode fiber. The cables on the right side of this photo are quad-link ribbon fiber for ultra-high density for inter-switch links. The optical signaling is 2.5GBaud, 8b/10b encoded.

16-Port Myrinet-2000 Switch



Myricom makes small Myrinet switches,
but we are best known for our large switches

Myrinet-2000 Switches for Large Clusters

512 Ports

in the Clos256+256 configuration pictured, 256 host ports plus 256 inter-switch ports, **or**

256 Ports

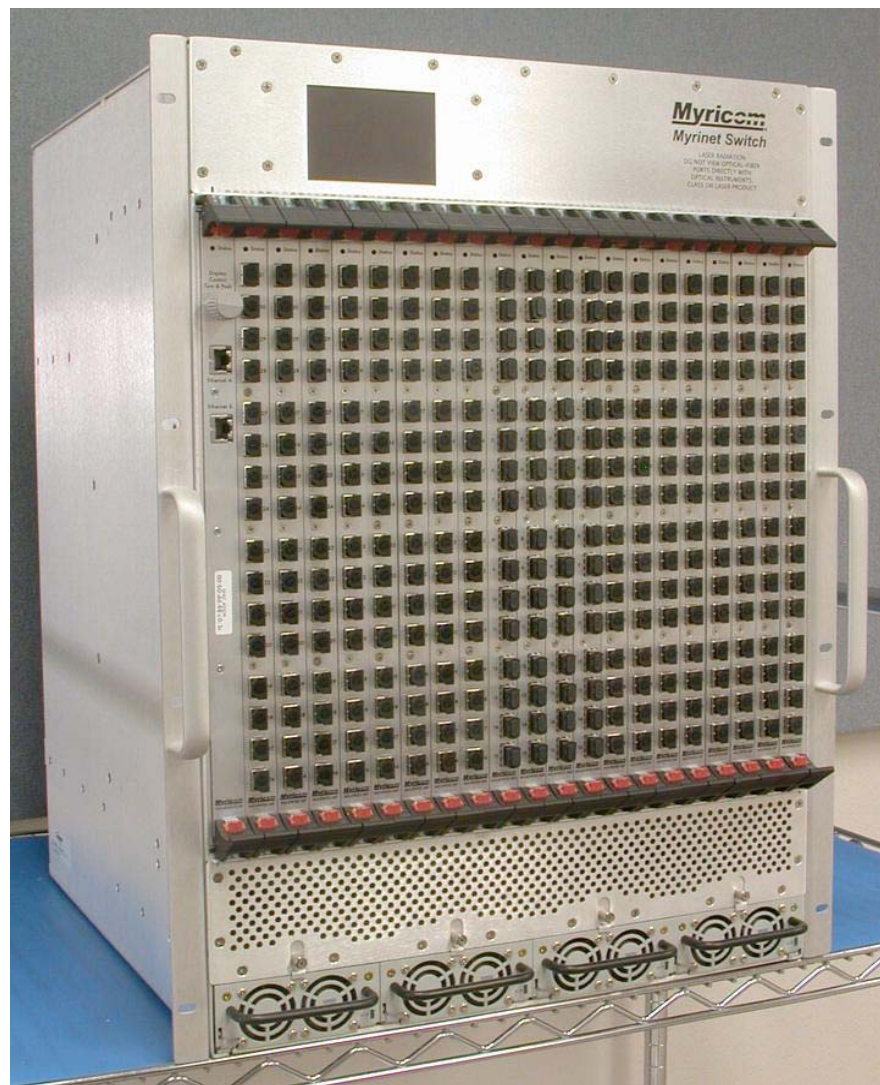
in the low-cost standalone Clos256 configuration, **or**

Up to 1280 Ports

in the Spine configuration used to connect Clos256+256 switches.

Full-bisection Clos networks.

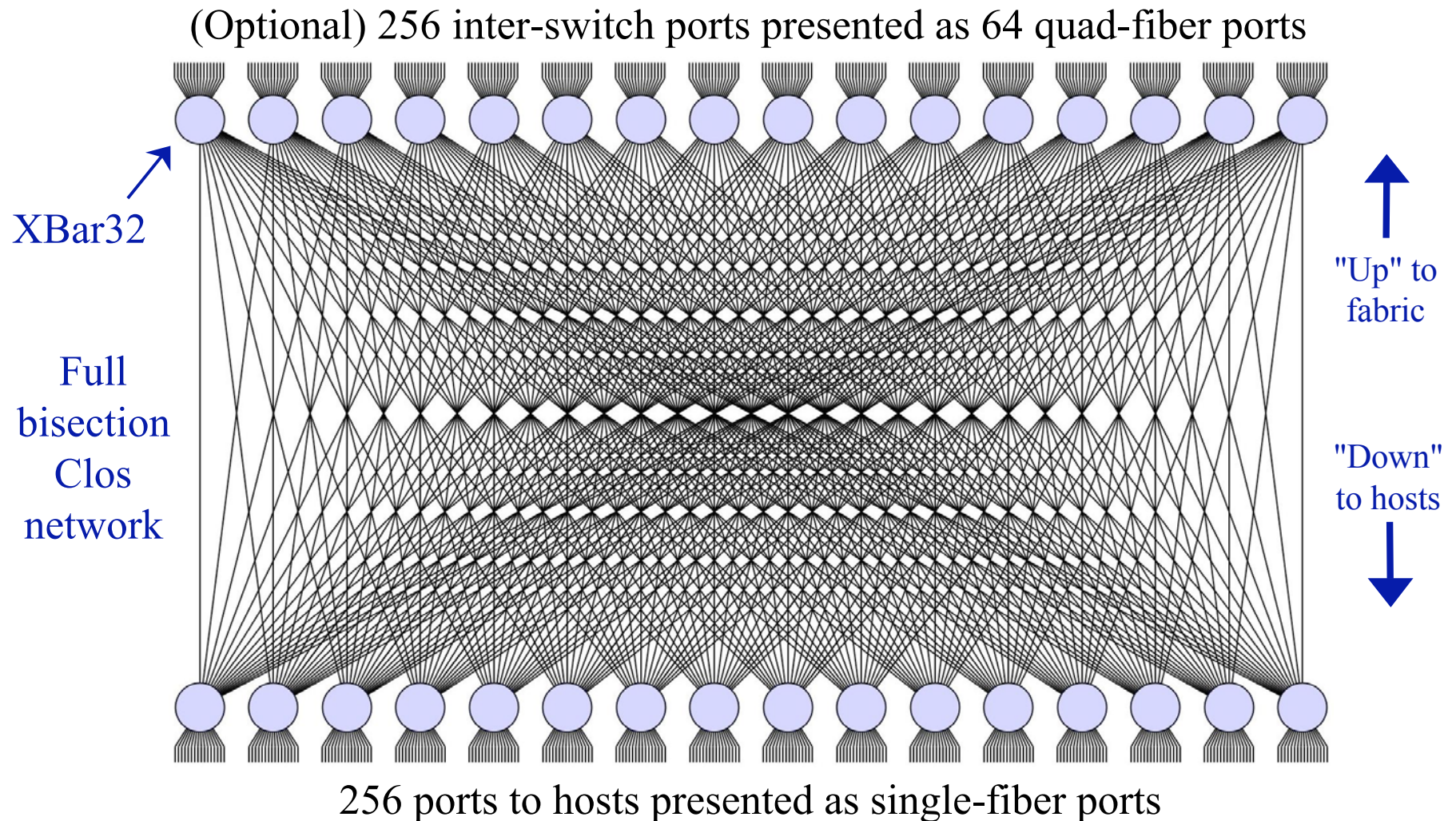
Extensive enterprise features.



www.myri.com

© 2005 Myricom, Inc.

Internal Clos Topology of the Clos256+256

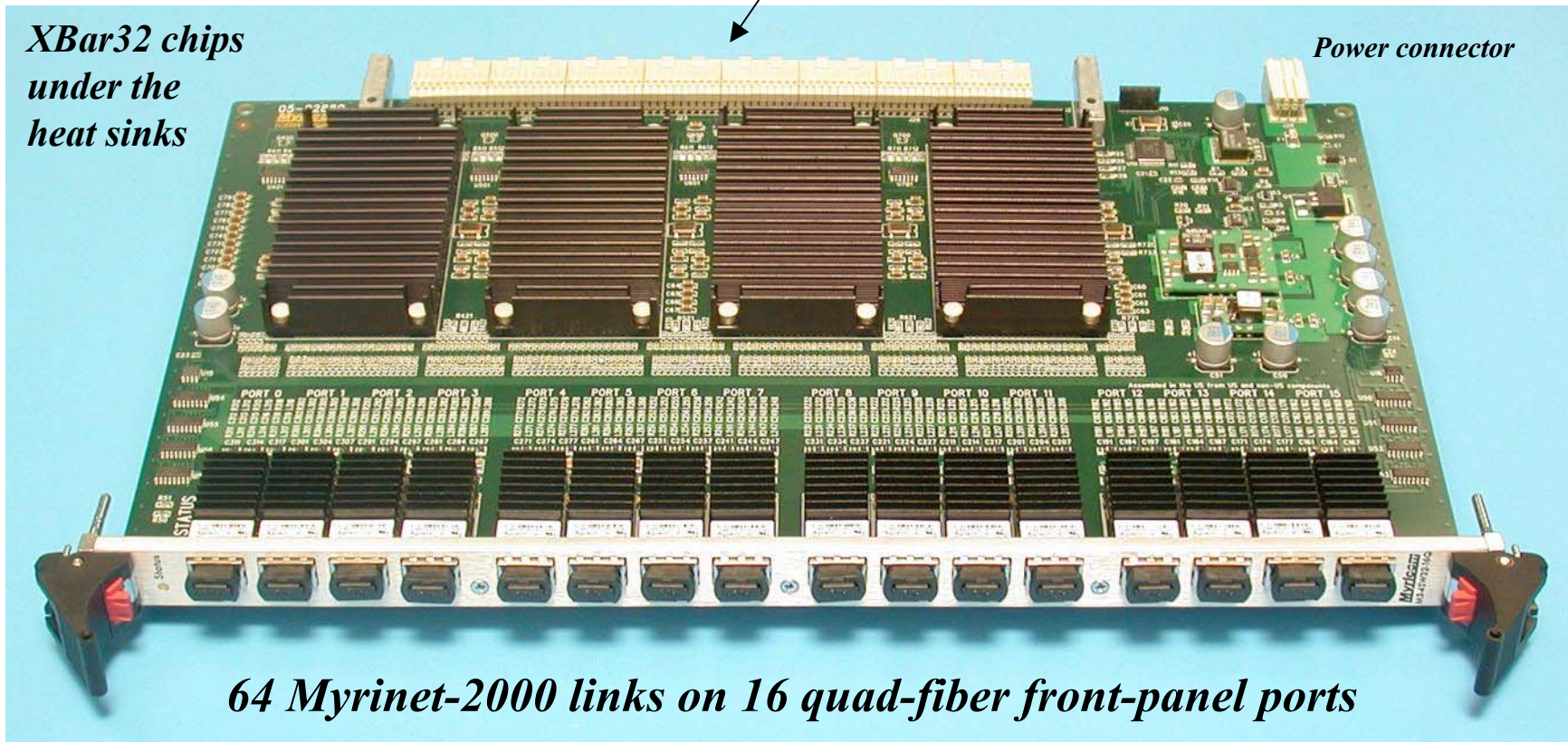


A glimpse of the technology inside these switches

64 Myrinet-2000 links on the high-speed-signal connectors to the backplane

*XBar32 chips
under the
heat sinks*

Power connector



64 Myrinet-2000 links on 16 quad-fiber front-panel ports

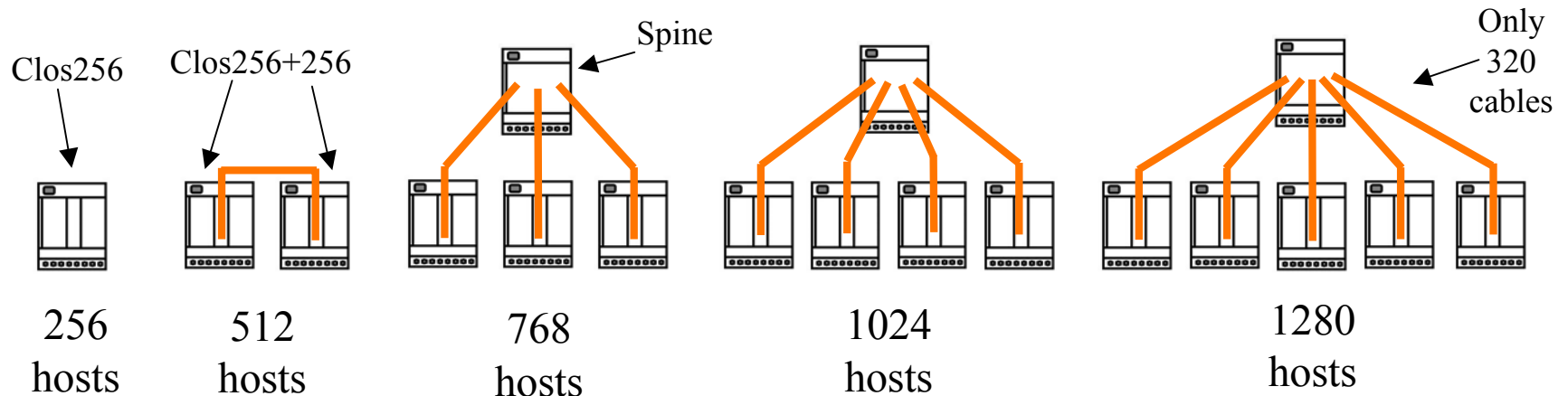
Very high switching density, e.g., much higher switching and port density than components for GbE switches, and on links that operate at twice the data rate.

Myricom®

www.myri.com

© 2005 Myricom, Inc.

Designed to Scale



- 1536 host ports requires 6 Clos256+256 and 2 partially populated Spine units
- ...
- 2560 host ports requires 10 Clos256+256 and 2 fully populated Spine units

All inter-switch cabling on quad-link ribbon fiber

A Myrinet Switch Network for 2560 Hosts

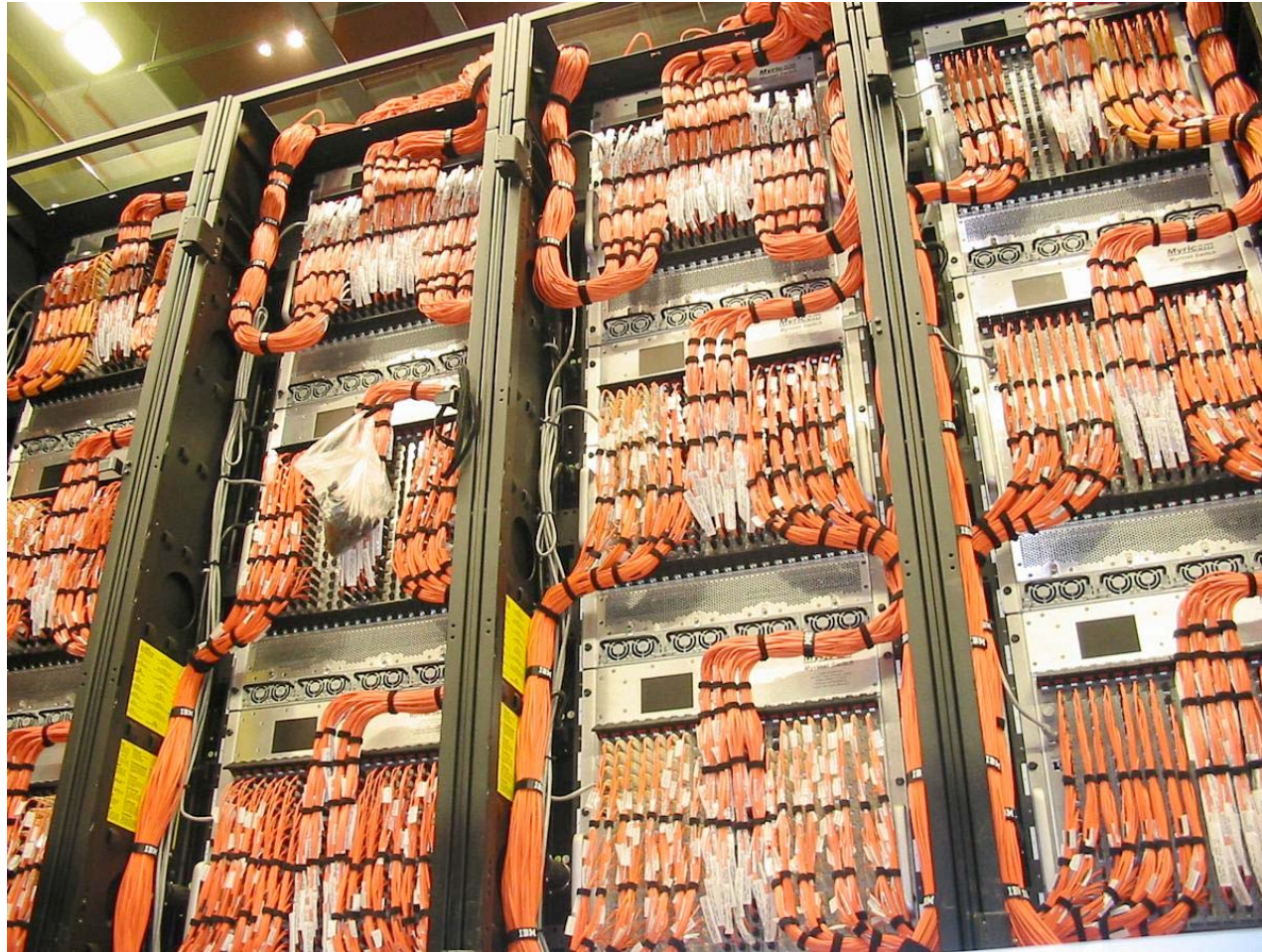


*Photo
courtesy
of IBM*

The switch network for the MareNostrum cluster at the Barcelona Supercomputing Center. MareNostrum was ranked #4 in the Nov-04 TOP500, and is the fastest cluster in the world.

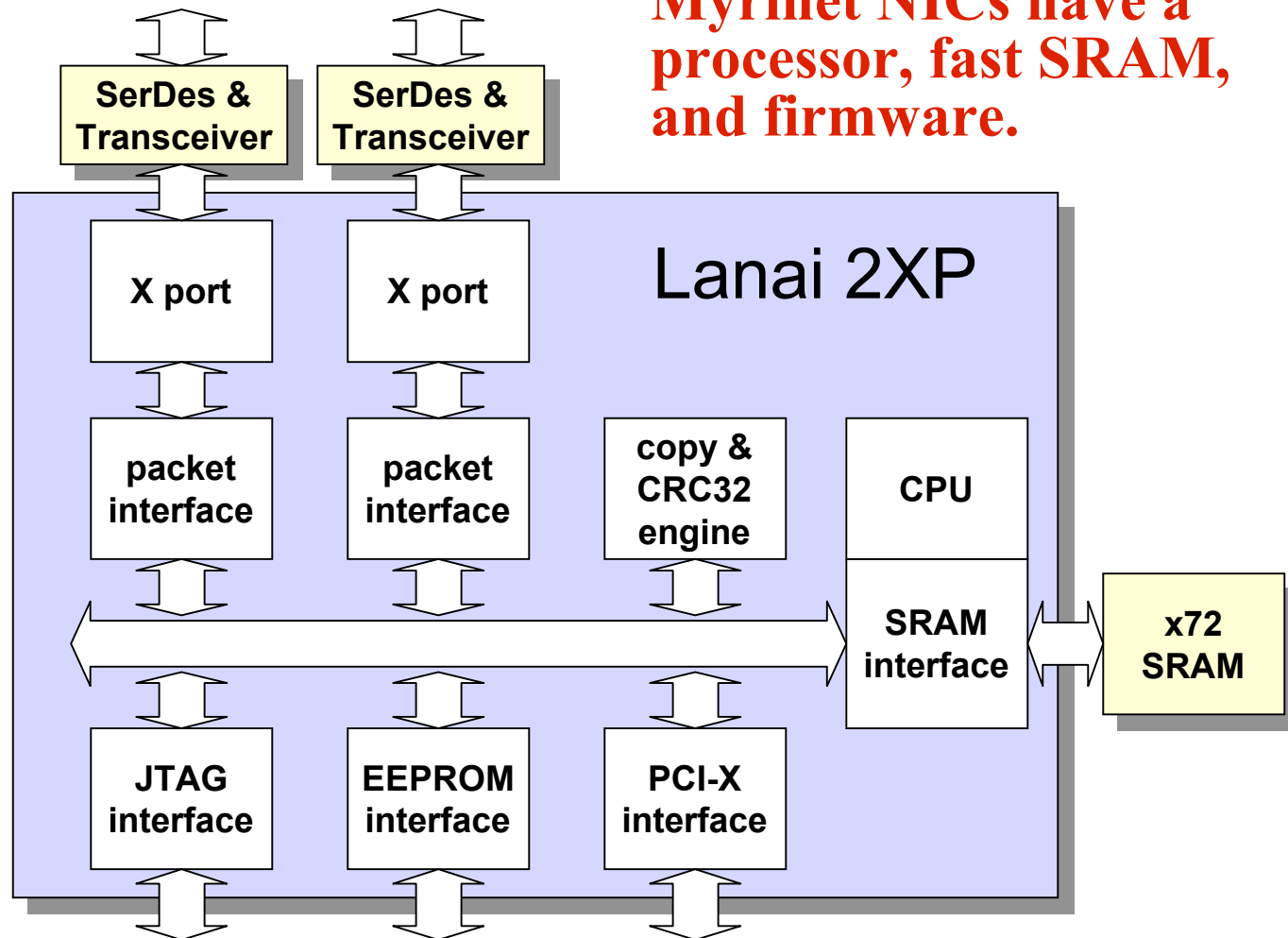
Myricom[®]

www.myri.com

© 2005 Myricom, Inc.

Myrinet NICs = Protocol Offload Engines

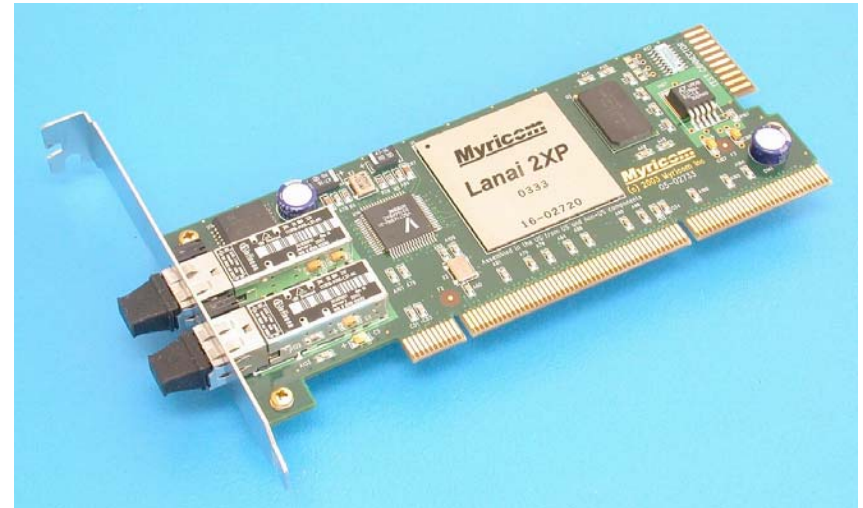
Myrinet NICs have a processor, fast SRAM, and firmware.



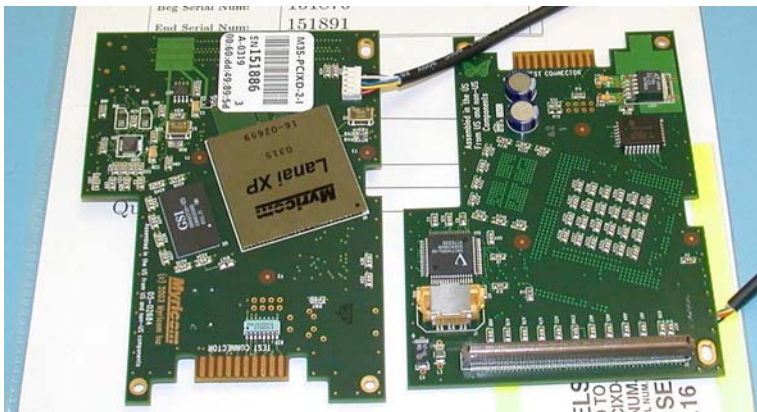
Current-production Myrinet-2000/PCI-X NICs



*One-port NICs:
"D card" (225MHz) & "F card" (333MHz)*



Two-port NIC: "E card" (333MHz)



IBM BladeCenter version of the D card

All PCI-X-series NICs are compatible at the network, software, and API levels. PCI-X performance on a dual-2.4GHz Xeon with the Serverworks chip set: 932 MB/s read, 1044 MB/s write. These NICs are self-initializing both for convenience and to allow diskless booting.

One more thing about Myrinet NICs...

A Myrinet NIC looks from the standpoint of installation exactly like an Ethernet NIC

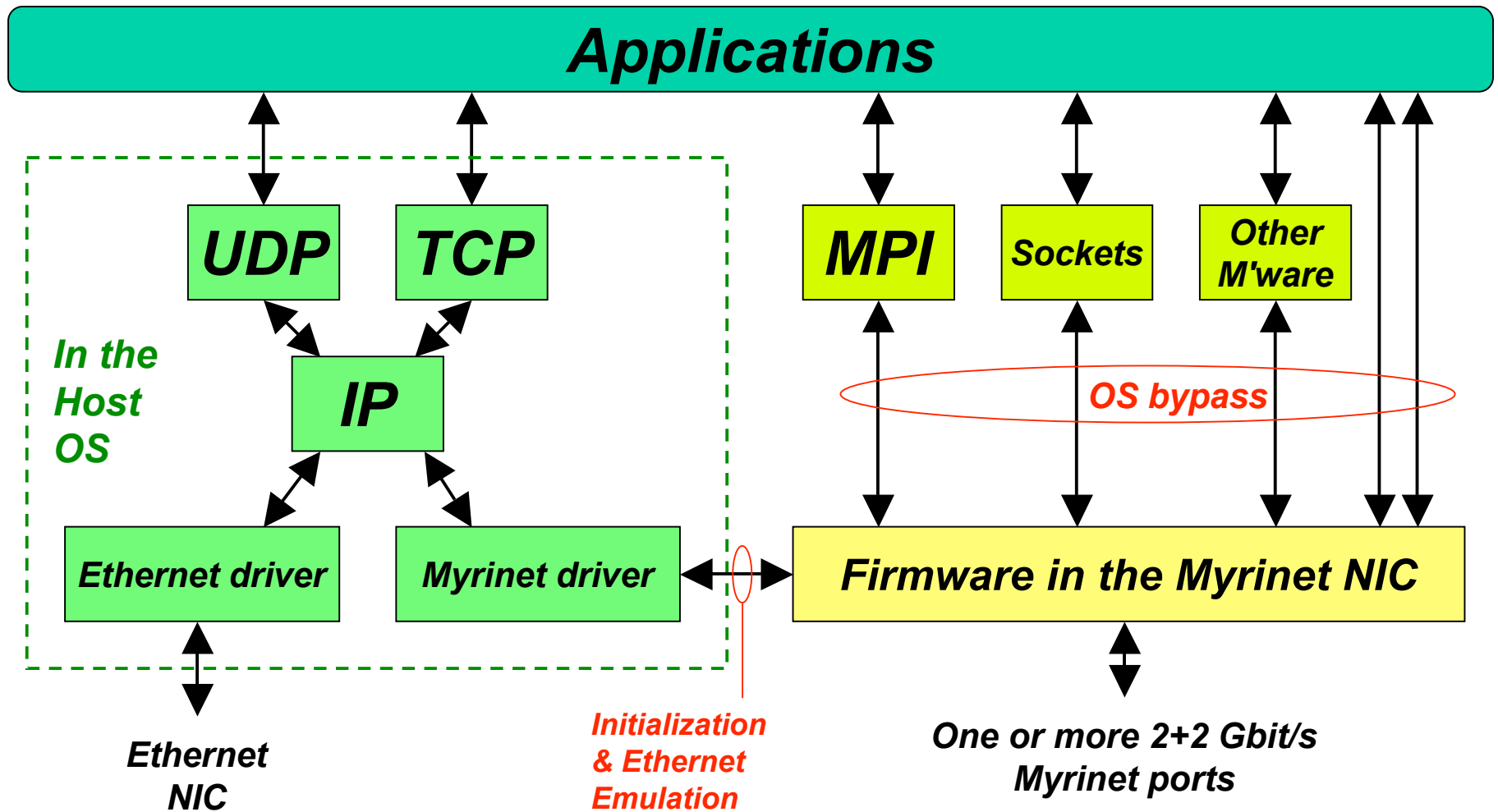


**An Ethernet
MAC address**

Myrinet NIC product label

The Myrinet device driver advertises itself to the host operating system as an Ethernet driver

Myrinet Software Interfaces



Choice of Myrinet Software Interfaces

*Applications need not be tailored for Myrinet.
Myricom provides the APIs the applications require.*

- **Low-level APIs**
 - GM 1 (legacy), GM 2 (current standard), MX (new), 3rd party (e.g., SCore)
- **TCP/IP & UDP/IP – *Commercial Applications***
 - Ethernet emulation, included in all GM and MX releases
 - 1.98 Gb/s (D or F cards) or 3.95 Gb/s (E cards) TCP/IP netperf on Linux (2.6.11 smp kernel)
- **MPI – *HPC Applications***
 - An implementation of Argonne MPICH directly over GM or MX.
 - Third-party MPI implementations over Myrinet are also available.
- **Sockets – *High-Performance TCP/IP Applications***
 - An implementation of UNIX or Windows sockets (or DCOM) over GM or MX. Completely transparent to application programs. Use the same binaries!
- **uDAPL and kDAPL – *Database Cluster Applications***
 - The new standard for distributed databases (Oracle, DB2)

Myrinet Express (MX)

- MX was designed based on earlier experience in writing MPICH-GM and VI-GM middleware
 - MX does generalized matching in the NIC firmware.
 - One of the keys to middleware performance
 - Perfect overlap of computation and communication under MPI.
 - Excellent MPI ping-pong latency

Myrinet-2000 NIC	D card (one port)	E card (two ports)	F card (one port)
MPICH-MX/MX unidirectional throughput	230 MB/s	475 MB/s	240 MB/s
MPICH-MX/MX bidirectional throughput	450 MB/s	840 MB/s	475 MB/s
MPICH-MX/MX latency	3.2μs	2.6μs	2.6μs

MX-2G between dual-Opteron hosts, including the latency through one switch

Adaptive Route Dispersion

- MX takes advantage of the multiple paths through large networks (*c.f.*, slides 6 & 8) to spread packet traffic
 - MX mapping provides multiple routes to each other host.
 - MX measures the time to inject each packet in order to sense contention. Brief flow-control backpressure indicates contention on the route.
 - MX changes route only when contention is sensed in the network
 - Automatically adapts to current traffic patterns.
 - Note: MX can receive packets out of order
 - But matching (message level) is always in-order.
- Eliminates "hot spots" in switch networks
 - Adapts the routes to the communication patterns of the application.
 - Extremely valuable for large switch networks.
- Only possible with source-routed networks

Myrinet-2000 Summary

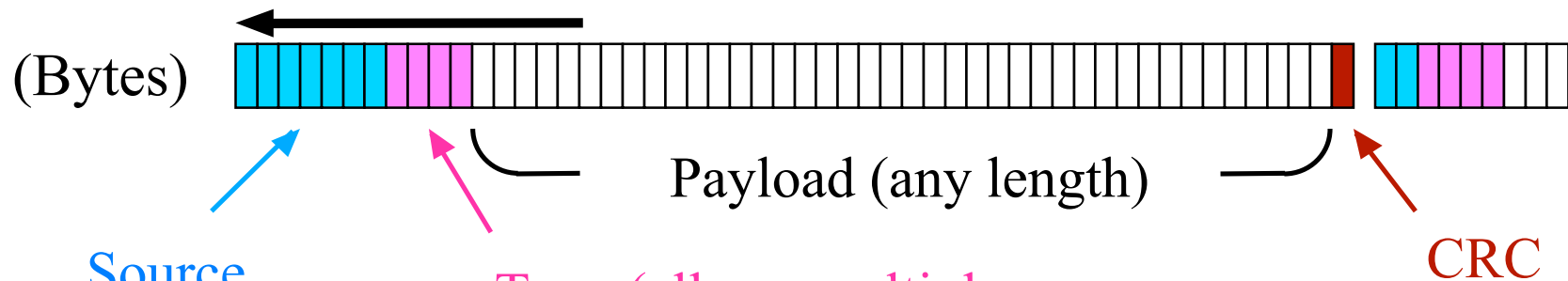
- **Low latency**
 - 2.6 μ s (MPI user level) for the fastest NICs, or
 - 3.2 μ s for the lowest-cost NICs
- **High data rate**
 - 2+2 Gb/s (250+250 MB/s) data rate links; user level is 95-99% of peak.
 - For higher data rates, use one or more dual-port NICs.
- **Very low host-CPU utilization**
 - Protocol processing is offloaded
 - $\log P < 0.3\mu$ s
- **Unlimited scalability**
 - Switch cost per host scales very well in the range $16 \leq N \leq 8192$
- **Multimode-fiber links** to 200m
 - Lightweight, small diameter, reliable
- **High Availability features**
 - Self-mapping, self-healing
 - Link-continuity monitoring
- **Data Integrity features**
 - Memory and bus parity
 - Link and packet-payload CRCs
- **Software drivers for almost all major platforms**
 - Download them from the Web
 - Open source
 - Low-level APIs + TCP|UDP/IP + MPI + VI + PVM + Sockets + DAPL
- **Hybrid Myrinet/Ethernet networks**

Architecture: “The Mirror Image of Ethernet”

- A network architecture, protocol, and **technology**
 - A descendant of packet communication and routing in MPPs (Massively Parallel Processors), but open.
 - ANSI Standard (ANSI/VITA 26-1998).
 - **The mirror image of Ethernet:** With Ethernet, the NICs are simple and the switches are complex. With Myrinet, processing power is concentrated in the hosts and NICs, allowing an elegant, streamlined, switching technology.
 - *Of course, Myrinet can do everything that Ethernet does. It just does it differently.*

Myrinet = ANSI/VITA 26-1998

Myrinet is defined at the Data-Link level (level 2 of the ISO reference model for computer networks) by its packet format and flow control. *Think of Myrinet as the simplest packet-switched network you could devise.*



Source
route
used
by the

Type (allows multiple
protocols on one
Myrinet)

<http://www.myri.com/open-specs/>

switches, which strip the bytes as they are used

There is flow-control and heartbeat monitoring on every link.

Why does Myrinet work so well for clusters? (1)

- The "Processor & Firmware in the NIC" Architecture
 - *Versus Ethernet or IB: Simple or RDMA NICs ...*
 - ... which depend upon the host to handle low-level network operations
 - Myrinet NICs are firmware-driven offload engines
 - The NIC processing supports OS-Bypass operation
 - Message operations without system calls, resulting in low latency and low-host-CPU overhead
 - The NIC processing provides type matching
 - Immediate, first-level demultiplexing of incoming messages, resulting in efficient use of IO bandwidth, host CPU, and host memory (no "RDMA window" memory use)
 - The NIC processing handles network protocols
 - Mapping, dispersive routing, reliability layer (reliable ordered delivery with acknowledgments), and exception handling.

Why does Myrinet work so well for clusters? (2)

- Source Routing

- *Versus Ethernet: Destination Routing, in which ...*
 - ... the switch must decide how to route a packet to a destination
 - and the switch is limited to information local to this switch
 - ... switching is typically store-and-forward
- Myrinet Switches are based on simple, high-degree, crossbar switches
 - Inexpensive, low latency, and highly scalable
 - The route is predetermined at the source: the switch just steers packets
 - The network end points have global information, and can be much "smarter" about routing (e.g., dispersive routing) than any single switch
 - Cut-through routing -- low latency even through many switch "hops"

- Note: Myrinet and Quadrics both employ these two techniques: a processor and firmware in the NIC, and source routing
 - The preferred and proven architecture for scalable HPC clusters

Myrinet Milestones – Technical Progress

- 1994: Myricom founded; first Myrinet-640 product shipments
- 1996: First Myrinet-1280 product shipments
- 1997: First cluster in the TOP500
 - Berkeley NOW, a Myrinet cluster of 100 SPARCs
- 1998: Myrinet becomes an ANSI Standard
- 2000: First Myrinet-2000 product shipments; GM-1 software
- 2001: All-fiber Myrinet-2000
- 2003: PCI-X NICs; GM-2 software
- 2004: Fastest TOP500 cluster: MareNostrum
 - Based on new Myrinet switches for large clusters
- 2005: Introduction of Myri-10G products & MX software
 - Dual-use products for 10-Gigabit Myrinet and 10-Gigabit Ethernet

Myri-10G

- **Convergence of 10-Gigabit Myrinet and 10-Gigabit Ethernet**
 - 4th-generation Myrinet (Myrinet-10G)
 - Any 10-Gigabit Ethernet Physical layer (10+10 Gb/s data rate links)
 - **Ethernet interoperability**: NICs are dual-protocol. Also, you can connect from Myrinet switch fabrics to IP networks or storage through simple protocol-conversion devices
 - Application-level compatibility with Myrinet-2000
 - Same architecture, same proven software, same APIs
- **Myrinet Express (MX) software**
 - MX-10G is the message-passing system for Myri-10G products
 - 10-Gigabit Ethernet software is derived from MX ethernet emulation
 - MX-2G is already released for Myrinet-2000 D, E, & F cards
 - Myricom software support always spans two generations of NICs

The Future of Myrinet-2000

We expect Myrinet-2000 products to continue to sell well through 2006, and to co-exist in the marketplace with the higher performance and higher priced Myri-10G products.

Myrinet-2000 is well positioned as a superior alternative to Gigabit Ethernet for clusters, while Myri-10G will offer performance and cost advantages over 10-Gigabit Ethernet for clusters.

Summary of 4th-Generation Myrinet (1)

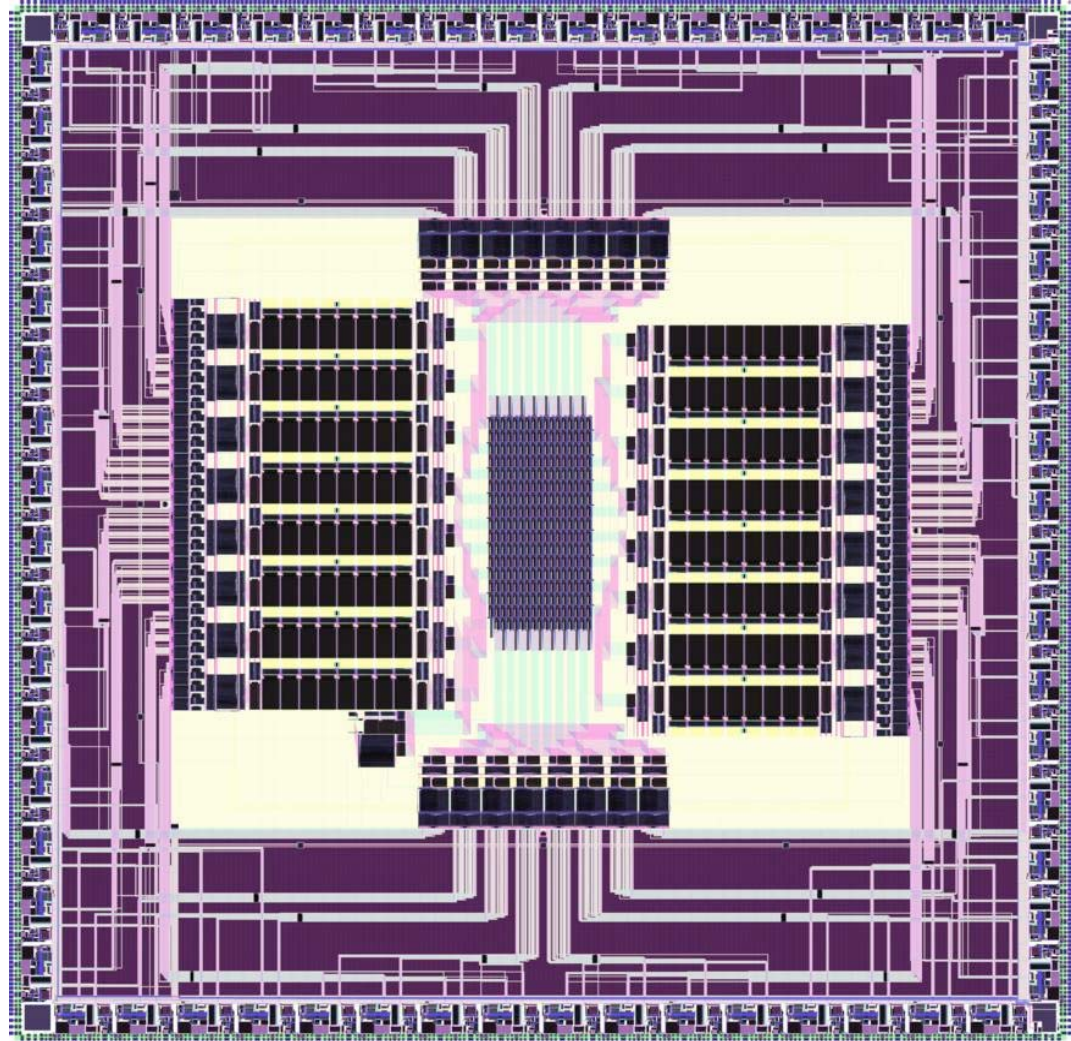
- **Links are 10-Gigabit Ethernet**
 - Any 10-Gigabit Ethernet PHY can be used for Myri-10G
 - The ports of Myricom's Myri-10G chips are XAUI, per IEEE 802.3ae
- **NICs have "Myri-10G" dual-protocol ports**
 - With bundled 10-Gigabit Ethernet software support (driver and firmware), the NIC operates as a protocol-offload 10-Gigabit Ethernet NIC
 - With optional Myrinet Express (MX-10G) software support, and when connected to a 10G Myrinet switch, the NIC operates in Myrinet mode
- **Software support is Myrinet Express (MX)**
 - Includes ethernet emulation (TCP/IP, UDP/IP) and a full suite of popular APIs, including MPICH-MX, Sockets-MX, and DAPL-MX
 - MPICH2-MX coming soon
 - Bundled driver+firmware for 10-Gigabit Ethernet operation is based on MX ethernet emulation, but simplified by omitting Myrinet mapping and protocols

Summary of 4th-Generation Myrinet (2)

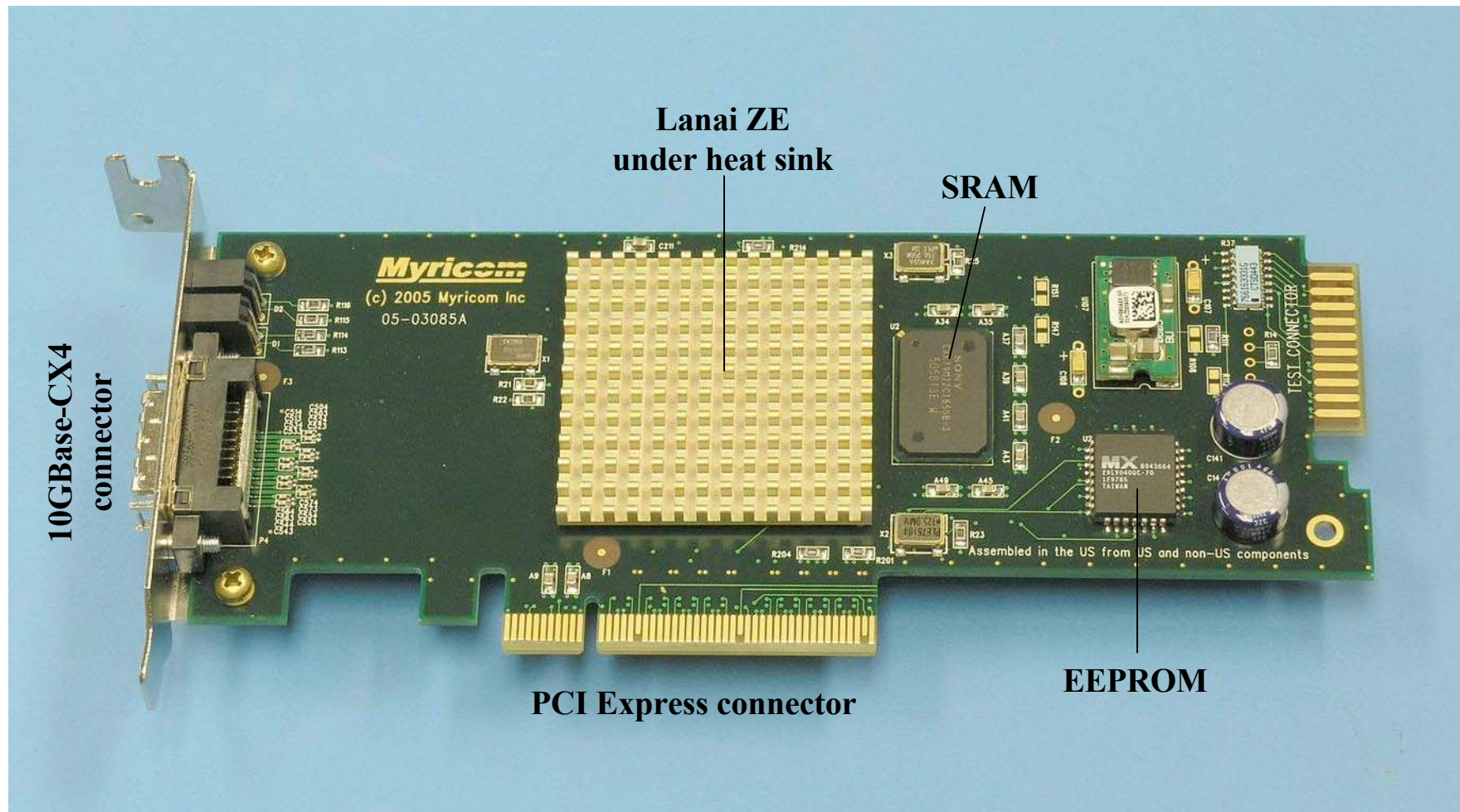
- **Switches are 10-Gigabit Myrinet**
 - Although we leverage 10-Gigabit Ethernet technology at the Physical level, we preserve the efficiency, simplicity, and scalability of Myrinet switching
 - Simple protocol conversion between 10G Myrinet switch fabrics and 10G Ethernet ports for connection to IP networks and storage
 - Myrinet-10G switches are similar except for data rates to Myrinet-2000 switches, and are based initially on a 16-port single-chip crossbar switch
- **Cables: choice of copper or fiber**
 - For low-latency applications, the initial choices are 10GBase-CX4 cables to 15m (lowest cost), and quad ribbon fiber cables to 200m
 - Additional 10-Gigabit Ethernet PHYs will be available later
- **Performance with the initial Myri-10G/PCI-Express NICs**
 - **Myrinet mode:** 2 μ s MPI latency with MPICH-MX, and 1.2 GBytes/s one-way data rate (Pallas benchmarks)
 - **10-Gigabit Ethernet mode, or under MX-10G ethernet emulation in Myrinet mode:** 9.6 Gbits/s TCP/IP rate (netperf benchmarks)

10-Gigabit Myrinet Switches

- Based on the 10G_XBar16 chip pictured
 - 16 XAUI ports
 - Power: ~15W
- Packaged similarly to Myrinet-2000 switches
- The same mapping-acceleration features as the Myrinet-2000 XBar32 chip



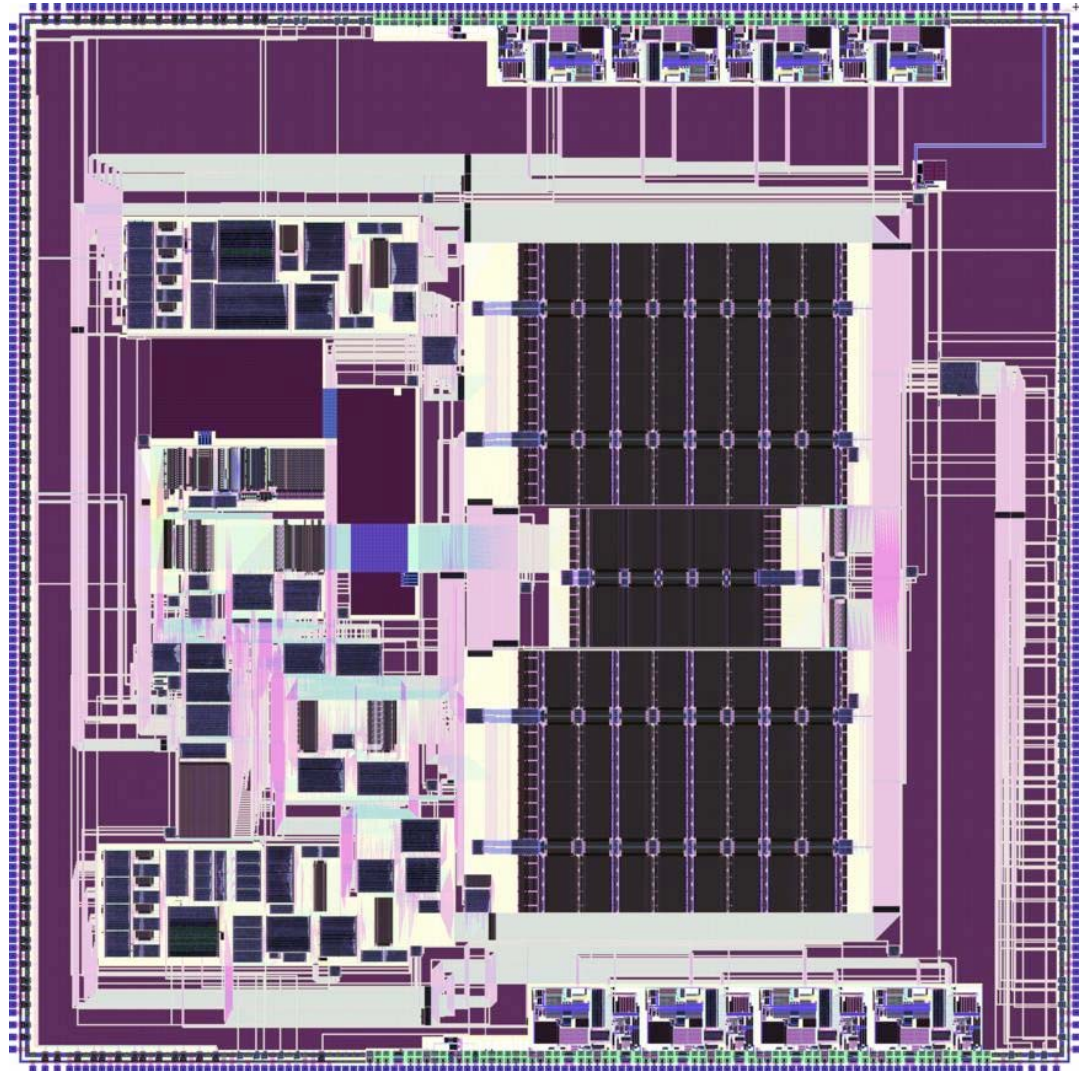
Myri-10G NIC (PCI Express)



Photograph of a software-development prototype. These powerful NICs are even more highly integrated than the Myrinet-2000 PCI-X NICs. The complexity is inside the Lanai-ZE chip.

The Lanai-ZE Chip

- PCI Express port
 - 4-lane or 8-lane
- Dual-protocol Myri-10G port (XAUI)
- High throughput
 - Internal packet buffers in addition to the external SRAM
- 313MHz processors and external memory
- Next: Lanai-ZH for HyperTransport



Questions?

LANL Lightning Cluster, 1408 dual Opterons,
entered TOP500 in Nov-2003 ranked #6,
8051 Gflops, 71.5% of peak.

